Nucleic Acids Research, 2004, Vol. 32, Database issue ***D471–D475***
DOI: 10.1093/nar/gkh123

FREP: a database of functional repeats in mouse cDNAs

Takeshi Nagashima[1], Hideo Matsuda[2], Diego G. Silva[3,4], Nikolai Petrovsky[3,4], RIKEN GER Group[5] and GSL Members[6], Akihiko Konagaya[7] and Christian Schönbach[1,*]

[1]Biomedical Knowledge Discovery Team, Bioinformatics Group, RIKEN Genomic Sciences Center (GSC), RIKEN Yokohama Institute, Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan, [2]Department of Bioinformatic Engineering, Graduate School of Information Science and Technology, Osaka University, Toyonaka, Osaka 560-8531, Japan, [3]Autoimmunity Research Unit, The Canberra Hospital, Woden ACT 2605, Australia, [4]John Curtin School of Medical Research, Australian National University, Canberra ACT 2601, Australia, [5]Genome Research Exploration (GER) Group, RIKEN Genomic Sciences Center (GSC), RIKEN Yokohama Institute, Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan, [6]Genome Science Laboratory, RIKEN, Hirosawa, Wako, Saitama 351-0198, Japan and [7]Bioinformatics Group, RIKEN Genomic Sciences Center (GSC), RIKEN Yokohama Institute, Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan

Received August 8, 2003; Revised and Accepted October 20, 2003

ABSTRACT

The FREP database (http://facts.gsc.riken.go.jp/FREP/) contains 31 396 RepeatMasker-identified non-redundant variant repeat sequences derived from 16 527 mouse cDNAs with protein-coding potential. The repeats were computationally associated with potential effects on transcriptional variation, translation, protein function or involvement in disease to identify Functional REPeats (FREPs). FREPs are defined by the (i) occurrence of exon–exon boundaries in repeats, (ii) presence of polyadenylation sites in 3′UTR-located repeats, (iii) effect on translation, (iv) position in the protein-coding region or protein domains or (v) conditional association with disease MeSH terms. Currently the database contains 9261 (29.5%) inferred FREPs derived from 6861 (41.5%) mouse cDNAs. Integrated evidence of the functional assignments and dynamically generated sequence similarity search results support the exploration and annotation of functional, ancestral or taxon-specific repeats. Keyword and pre-selected feature searches (e.g. coding sequence–repeat or splice site–repeat relations) support intuitive database querying as well as the retrieval of repeat sequences. Integrated sequence search and alignment tools allow the analysis of known or identification of new functional repeat candidates. FREP is a unique resource for illuminating the role of transposons and repetitive sequences in shaping the coding part of the mouse transcriptome and for selecting the appropriate experimental model to study diseases with suspected repeat etiology contributions.

INTRODUCTION

Repetitive DNA sequences include simple repeats and transposon-derived repeats. The latter are also called interspersed repeats and occupy 46% of the human (1) and 38% of the mouse (2) genomes. Transposon-derived repeats fall into two categories depending on their expansion mechanism. Long interspersed elements (LINEs), short interspersed elements (SINEs) and long terminal repeats (LTRs) transpose through an RNA intermediate, whereas DNA transposons (e.g. Mariner) are transposed as DNA (3). Simple repeat sequences (4) with repeating units of 1–13 nucleotides (e.g. dinucleotide, triplet repeats and microsatellites) or 14–500 nucleotides (minisatellites) are probably expanded through errors in DNA replication (5). The distribution, frequency and diversification of repeats in human and mouse differ. For example, mouse contains a higher number of recent transposons that diversify more rapidly than in human. The differences also extend to simple DNA repeats in protein-coding regions. Most neurodegenerative disease-associated poly-glutamine translated CAG triplet repeats and myotonic dystrophy-associated CTG triplet repeats of the untranslated region (UTR) are longer and appear to be more stable in normal mice than human (6). The taxon-specific differences in repeat evolution complicate the inference of specific functions associated with a repeat and their extrapolation to other species.

Several databases support the identification and classification of repeat sequences at the DNA and protein levels. For example, RepBase (7) is a general reference nucleotide sequence repeat repository that is frequently used with RepeatMasker (http://repeatmasker.genome.washington.edu/) to mask or classify repeats. Repeating protein sequences, some of them originating from simple DNA repeats, are classified and compiled by InterPro (8) and Swiss-Prot (9). Other species-specifc repeat databases are often part of larger computational genome annotations (10) and DNA typing,

*To whom correspondence should be addressed. Tel: +81 45 503 9303; Fax: +81 45 503 9552; Email: schoen@gsc.riken.jp
RIKEN GER Group and Genome Science Laoratory (GSL) Members are: Takeya Kasukawa, Takahiro Arakawa, Piero Carninci, Jun Kawai and Yoshihide Hayashizaki

microsatellite mapping (11) or amplification typing of LINE active subfamilies (12). Associations of repeat sequences with inherited disease genes are reported in the Repeat Sequence Database (RSDB) (13). However its scope is restricted to human tandem repeats in genes with Online Mendelian Inheritance in Man (OMIM) (14) records.

Despite the value of these databases the extent and role of repetitive sequences in RNA processing, transcript termination, splicing, protein functions, protein domains or disease etiology remain largely unexplored. Since the FANTOM2 (Functional Annotation of Mouse) project (15) provided us with a representative transcript and protein set derived from 60 770 RIKEN full-length cDNAs and 44 106 public mouse (*Mus musculus*) mRNA sequences we conducted a functional repeat analysis on 74 031 mouse transcripts with protein-coding potential and integrated the results with computationally inferred functional repeat associations and cross-species comparisons into the Functional Repeat (FREP) database.

DESCRIPTION

The FREP database stores the data of 40 701 repeats that were identifed with RepeatMasker in 21 808 of 74 031 cDNAs with coding sequence (CDS) information. These cDNAs correspond to 7027 representative transcripts and 14 229 representative repeats (RRS). Details of the sequence sources and construction of RRS, variant repeat (VRS) and variant repeat transcript (VRTS) sets are given in the FREP documentation (http://facts.gsc.riken.go.jp/FREP/doc/documentation.html).

In brief, cDNAs that cluster may show variations in repeat length, substitutions among the same type of repeats or deletion/insertion of a repeat in comparison with the representative transcript and its representative repeats. The variant cDNAs and their corresponding non-redundant variant repeats were assigned to the VRTS and VRS, respectively. The VRS comprises 31 396 repeat sequences derived from 16 527 VRTS members. Here, we use VRS and VRTS to describe the content of the FREP database. FREPs, as defined in the abstract, comprise 29.5% (9261) of identified VRS repeats.

Data and content

About 4.5% (1401) of variant repeats in 7.6% (1253) VRTS members were aligned by Sim4 (16) to splice junctions in genomic regions extracted with BLAT (17) and appear to be spliced repeats. Notably, 17% (239) of them contribute to alternative splicing in 229 VRTS corresponding to 4.6% (219 of 4750) of all mouse splice variant clusters (18). 3′UTR-located repeats containing poly(A) signals of various strength may generate alternative poly(A) site usage and therefore differential transcription termination and RNA processing (19). Of conserved AATAAA poly(A) sites in 1884 cDNAs, 2902 are of repeat origin. About 9.2% (2881) of VRS are located in the protein-coding region of 2331 VRTS members. Of these, 460 translated repeats contribute to InterPro motifs, including protein repeats. The remaining protein-coding repeats comprise a source of mouse-specific variations and potential new motif candidates. By Sim4, 2818 VRTS members including 3364 variant repeats aligned to both mouse ($\geq$95% identity over $\geq$95% length) and human ($\geq$60% identity over $\geq$60% length) genomic regions. The repeats corresponding to these orthologs are therefore most likely ancestral repeats. Of the orthologous VRTS, 8.2% (235) comprise 271 inferred human disease-associated VRS repeats. Detailed dynamically generated statistics of repeat-clone distributions and functional associations by repeat classes and families are available at the FREP website (http://facts.gsc.riken.go.jp/FREP/statistics.html).

Repeats and associated data are stored in 21 relational database tables which are used to generate FREP summary lists and full reports. A FREP report (Fig. 1) consists of table-formatted basic clone information, a repeat summary with graphical evidence display and functional repeat classification, external links and other functional information.

Basic information

The table 'Basic Information' contains brief descriptions of the data source and cDNA such as the gene name and symbol, hyperlinked database source and representative transcript. A ClustalW (20) sequence alignment of the cDNA sequence that clusters with members of the representative transcript is hyperlinked.

Repeat summary

The repeat summary integrates RepeatMasker-derived (A. F. A. Smit and P. Green, http//ftp.genome.washington.edu/RM/RepeatMasker.html) repeat classifications, length, similarity to RepBase reference repeats [Smith–Waterman (SW) score], G+C content and functional associations into a simple overview that supports quick decision-making whether to explore the evidence and FREPs. Repeat sequence alignments and excised and/or translated repeat sequences are accessible through hyperlinks. Inferred FREPs are summarized by keywords signifying potential involvement in, or effect on splicing, translation, protein function, polyadenylation or disease associations. Links to Sim4 cDNA alignment results (table and graphical view) for mouse and human genomes provide evidence for repeats with splice junctions and for orthologous repeats. BLASTN (21) results (E-value $\leq$ e-50) of primate and rodent (except *M.musculus*) mRNA sequences are summarized by species names and keywords indicating repeat or sequence conservation.

Graphical summary

Each FREP report contains a dynamically generated colored graphical view that visualizes repeats with position and orientation information in relation to the cDNA and length, G+C content, CDS, domain(s), poly(A) signals in the 3′UTR region, chromosome assignment and exon–exon boundaries

FREP classification

The keyword-based FREP classification of repeats consists of repeat–CDS relationship, CDS problem information and inferred functional assignments: protein function, domain, splicing, poly(A) site and disease association. Evidence for each assignment is shown in the graphical summary or provided in hyperlinked records of the genomic exon alignments (splicing), MouSDB (18) variant exon mapping (alternative splicing) and color-coded repeat sequences [splice junctions, domain, poly(A) sites].

Repeat–disease associations were computationally assigned on three conditions: (i) the repeat is conserved in human and mouse genome mapping results, (ii) OMIM Morbidmap and

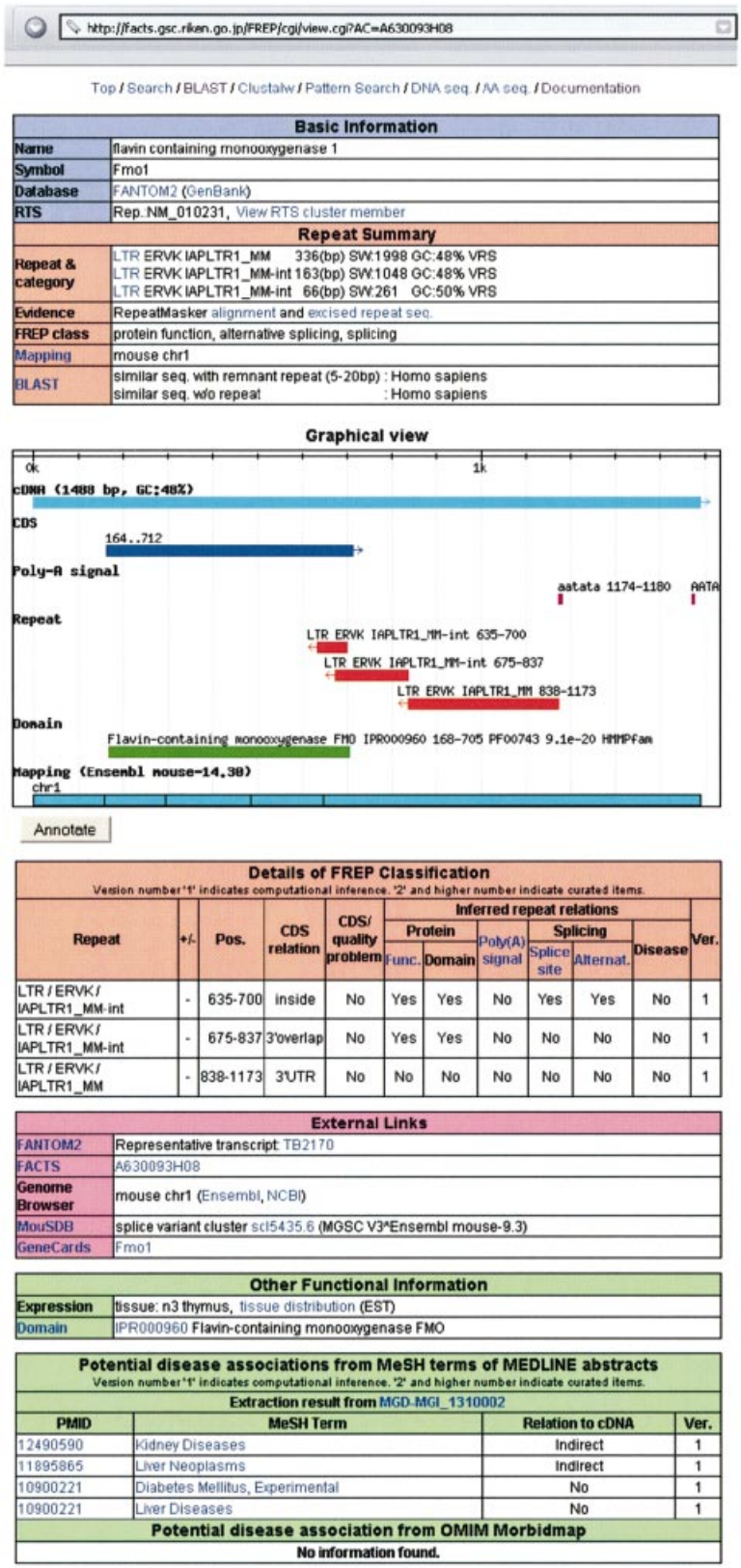

Figure 1. FREP report for flavin-containing monooxygenase 1 (AK042457). Repeat-relevant information is shown in the red-colored tables and the graphical view. *fmo1 c*ontains on the complementary strand a retrovirus-like IAP element of the LTR ERV-K family. The IAP supplied an alternative splice site which generated a splice variant with a 38 nucleotide repeat–CDS overlap, causing a premature stop codon. This mouse *fmo1* variant might be translated to a shortened, catalytically altered or inactive protein that may affect drug metabolism.

MEDLINE queries with gene name or symbol yielded an OMIM title or a disease MeSH term in the MH field of the retrieved abstracts and (iii) one of the other functional associations is positive. The combination of three conditions significantly reduces the number of potential false positive repeat–disease associations that would arise from the presence of only a gene name in the OMIM title. If the OMIM or MEDLINE query did not produce any results because the gene name was uninformative (e.g. hypothetical protein) but the repeat was conserved in human and fulfilled the third condition, the term 'uncertain' was assigned. Details of the disease MeSH and OMIM extraction methods (22) are explained in the FREP documentation (http://facts.gsc.riken.go.jp/FREP/doc/documentation.html).

External links

For data interpretation support and update-related issues we integrated links to FANTOM2, FACTS (22) and MouSDB splice variant cluster databases. Links to mouse and human genome maps of ENSEMBL and NCBI have been included as a reference to the most recent genome map and for FREP cDNAs that did not pass our mapping thresholds.The link to GeneCards (23) using the gene symbol and alternative names supplements FREP functional information with human tissue expression data and non-MeSH-based human disease information to aid the interpretation of repeats assumed to be involved in disease etiology.

Other functional information

Other functionally relevant information includes protein domain names extracted from InterProScan (24) results. Tissue information from cDNA library source annotations and BLASTN searches of mouse ESTs are presented in a hyperlinked table to aid contextual interpretation and candidate selection for gene expression profiling.

FUNCTIONALITIES

Keyword search and retrieval with pre-defined conditions

FREP integrates several tools for searching and retrieving data by keywords (gene name, gene symbol or disease MeSH terms), accessions, repeat and pre-defined conditions, which can be selected in any combination. The criteria include: repeat class, family type, length, repeat–CDS relations, functional associations, mouse or human genome mapping and chromosomal location, repeat conservation, tissue distribution. The search results consist of a dated list including the submitted query values, accessions, gene names and basic information on the repeats. Complete FREP reports of the retrieved entries are accessible through the hyperlinked accesssions. FASTA-formatted repeat sequences can be retrieved by repeat family, length and SW score of the RepeatMasker output.

Sequence searches and alignments

We integrated in the FREP database four sequence analysis tools. The ClustalW sequence alignment tool enables users to align repeat sequences, analyse variations among repeats of one family and obtain a phylogenetic view from the TreePlot (O. Langella, CNRS UPR 9034) output. BLASTN or TBLASTN search interfaces permit sequence similarity searches against the FREP repeat sequence, FREP cDNA and GenBank-derived rodent (without *M.musculus*), primate and human mRNA sequence datasets. Since the FREP cDNA BLAST database search is enhanced with a color-coded display of the repeat type and position, as well as links to the corresponding FREP reports, the absence of repeats or repeat variations are easily detected. Retrieval of exact or variable tandem repeats and motif sequence identification is facilitated by the modified pattern_find tool (K. Hofmann, Swiss Institute for Experimental Cancer Research). Accessions in the output are hyperlinked to either FREP or GenBank entries.

Annotation

Computationally inferred functions enhance the biological interpretation of repeat-containing clones. However, they may constitute a source of error propagation if downloaded and integrated into a curated database. We therefore implemented an annotation interface that is accessible upon registration as annotator. Inferred FREP classifications, including disease MeSH and OMIM associations, can be confirmed, negated or re-assigned and commented on. Computational inferred functional repeat associations are flagged with the version number '1' whereas '2' or higher numbers indicate human curated associations. Annotated records are automatically updated and available for downloading in tab-delimited format.

DATABASE ACCESS

FREP is hosted on a 2CPU/2Gb memory LINUX computer, which is accessible at the URL http://facts.gsc.riken.go.jp/FREP/. The programs for functional classification and graphical display were written in Perl and are available upon request. Data are stored in a relational database (PostgreSQL) accessible through Graphical User Interface forms. Database search and result display functions were built with a combination of SQL commands and Perl programs. All the FREP sequence data can be downloaded as zip- and bz2-compressed files.

FUTURE DIRECTIONS

At present FREP is limited to RepeatMasker-identified repeats in mouse cDNAs with coding potential. Future updates will include data obtained with other repeat-finding programs as well as repeats of the increasingly important non-coding transcripts and regulatory genomic regions of mouse and human.

ACKNOWLEDGEMENTS

We thank Mihaela Zavolan (Rockefeller University) for providing us with the variant exon information. The short-term stay of D.G.S. at the RIKEN Genomic Sciences Center was sponsored by a scholarship from the Novartis Foundation, London, UK. This study has been supported by a Research Grant for the RIKEN GER Project and a Research Grant for the National Project on Protein Structural and Functional Analysis from MEXT, the Japanese Government to Y.H.

REFERENCES

1. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
2. Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M., An,P. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
3. Prak,E.T. and Kazazian,H.H.,Jr (2000) Mobile elements and the human genome. *Nature Rev. Genet.*, **1**, 134–144.
4. Epplen,J.T., Maueler,W. and Santos,E.J. (1998) On GATAGATA and other 'junk' in the barren stretch of genomic desert. *Cytogenet. Cell Genet.*, **80**, 75–82.
5. Richards,R.I. and Sutherland,G.R. (1994) Simple tandem repeats are not replicated simply. *Nature Genet.*, **6**, 114–116.
6. King,B.L., Sirugo,G., Nadeau,J.H., Hudson,T.J., Kidd,K.K., Kacinski,B.M. and Schalling,M. (1998) Long CAG/CTG repeats in mice. *Mamm. Genome*, **9**, 392–393.
7. Jurka,J. (2000) Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.*, **16**, 418–420.
8. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.
9. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
10. Yuan,Q., Ouyang,S., Liu,J., Suh,B., Cheung,F., Sultana,R., Lee,D., Quackenbush,J. and Buell,C.R. (2003) The TIGR rice genome annotation resource: annotating the rice genome and creating resources for plant biologists. *Nucleic Acids Res.*, **31**, 229–233.
11. Ruitberg,C.M., Reeder,D.J. and Butler,J.M. (2001) STRBase: a short tandem repeat DNA database for the human identity testing community. *Nucleic Acids Res.*, **29**, 320–322.
12. Badge,R.M., Alisch,R.S. and Moran,J.V. (2003) ATLAS: a system to selectively identify human-specific L1 insertions. *Am. J. Hum. Genet.*, **72**, 823–838.
13. Horng,J.T., Huang,H.D., Jin,M.H., Wu,L.C. and Huang,S.L. (2002) The repetitive sequence database and mining putative regulatory elements in gene promoter regions. *J. Comput. Biol.*, **9**, 621–640.
14. Hamosh,A., Scott,A.F., Amberger,J., Bocchini,C., Valle,D. and McKusick,V.A. (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **30**, 52–55.
15. Okazaki,Y., Furuno,M., Kasukawa,T., Adachi,J., Bono,H., Kondo,S., Nikaido,I., Osato,N., Saito,R., Suzuki,H. *et al.* (2002) Analysis of the mouse transcriptome based on functional annotation of 60 770 full-length cDNAs. *Nature*, **420**, 563–573.
16. Florea,L., Hartzell,G., Zhang,Z., Rubin,G.M. and Miller,W.A. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.*, **8**, 967–974.
17. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
18. Zavolan,M., Kondo,S., Schönbach,C., Adachi,J., Hume,D.A., RIKEN GER Group and GSL Members, Hayashizaki,Y. and Gaasterland,T. (2003) Impact of alternative initiation, splicing and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. *Genome Res.*, **13**, 1290–1300.
19. Edwalds-Gilbert,G., Veraldi,K.L. and Milcarek,C. (1997) Alternative poly(A) site selection in complex transcription units: means to an end? *Nucleic Acids Res.*, **25**, 2547–2561.
20. Chenna,R., Sugawara,H., Koike,T., Lopez,R., Gibson,T.J., Higgins,D.G. and Thompson,J.D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.*, **31**, 3497–3500.
21. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
22. Nagashima,T., Silva,D.G., Petrovsky,N., Socha,L.A., Suzuki,H., Saito,R., Kasukawa,T., Kurochkin,I.V., Konagaya,A. and Schönbach,C. (2003) Inferring higher functional information for RIKEN mouse full-length cDNA clones with FACTS. *Genome Res.*, **13**, 1520–1533.
23. Safran,M., Solomon,I., Shmueli,O., Lapidot,M., Shen-Orr,S., Adato,A., Ben-Dor,U., Esterman,N., Rosen,N., Peter,I. *et al.* (2002) GeneCards 2002: towards a complete, object-oriented, human gene compendium. *Bioinformatics*, **18**, 1542–1543.
24. Zdobnov,E.M. and Apweiler,R. (2001) InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.